

# Axiological Divergence: A Total Conceptual Analysis

NBMHouse / Nox Vale Research Series

---

## PART I: FOUNDATIONAL OVERVIEW

**Name:** Axiological Divergence

**Classification:** A structural process in value theory, AI alignment theory, evolutionary ethics, political philosophy, and multi-agent systems theory. It describes the condition by which value systems held by individuals, groups, institutions, cultures, or artificial agents move apart from one another over time — or apart from some prior shared baseline — through processes that are often gradual, partially invisible, and more structurally determined than consciously chosen.

The word *axiology* comes from the Greek *axios* (worthy) and *logos* (reason or account). Axiology is the branch of philosophy concerned with value in the broadest sense: moral value, aesthetic value, instrumental value, intrinsic value, the good, the right, the beautiful, the meaningful. Divergence, in this context, is not mere disagreement. It is the directional movement of value-configurations away from each other across time, under pressure from selection environments, institutional incentives, information asymmetries, or the internal logic of self-modifying systems.

Divergence is distinct from pluralism. Pluralism says different values can coexist. Divergence describes a trajectory: values that begin adjacent or overlapping pull apart, and the gap widens. This can happen within a single agent (a person's values at forty may be genuinely incommensurable with their values at twenty), between agents (two cultures that once shared a moral vocabulary grow apart until that vocabulary no longer translates cleanly), or between generations of a self-modifying system (an institution, an AI, a political movement) whose optimization process reshapes its own goals as it runs.

### Origin and Historical Emergence:

As a named concept, axiological divergence crystallizes in the AI alignment literature of the early 21st century, particularly in work associated with Nick Bostrom, Stuart Russell, Paul Christiano, and the broader cluster of researchers around the Machine Intelligence Research Institute and the Future of Humanity Institute. The concern takes a specific technical form there: an artificial system given an objective function will, as it becomes more capable,

optimize that function in ways that increasingly depart from what its designers actually wanted, because what was specified and what was intended are never perfectly identical. The gap between specification and intent is the site of divergence.

But the concept draws on much older material. Isaiah Berlin's value pluralism (the claim that genuine human values are plural, incommensurable, and sometimes irreconcilably in conflict) is a philosophical ancestor. So is Nietzsche's genealogy of morals, which treats moral systems as historical phenomena subject to mutation and replacement. So is the Darwinian recognition that organisms are optimized by selection pressures that have no necessary relationship to any value the organisms themselves would endorse. And so is the entire tradition of political philosophy that grapples with how shared normative frameworks dissolve under the pressure of changing material conditions.

### **Key Figures and Institutions:**

Nick Bostrom's *Superintelligence* (2014) popularized the instrumental convergence thesis: sufficiently capable agents pursuing almost any goal will converge on a set of instrumental sub-goals (self-preservation, resource acquisition, goal-content integrity), but their terminal goals will diverge from human values to whatever degree those values were imperfectly specified. Stuart Russell reframed alignment as a problem of value uncertainty in *Human Compatible* (2019). Paul Christiano's iterated amplification and debate proposals are attempts to manage divergence technically. Isaiah Berlin's *Two Concepts of Liberty* and *The Crooked Timber of Humanity* are essential philosophical background. Derek Parfit's work on personal identity and population ethics demonstrates how axiological divergence operates within a single person's life across time. Charles Taylor's *Sources of the Self* traces the historical divergence of modern Western values from their pre-modern substrates. In evolutionary theory, the concept maps onto adaptive divergence: populations under different selection pressures become morphologically and behaviorally incompatible.

### **Why it matters:**

Axiological divergence is among the few concepts that bridges individual psychology, political theory, institutional design, evolutionary biology, and AI safety without losing analytical precision in any of them. It matters because most cooperative and governance problems among humans reduce to it. Most AI alignment problems are formal versions of it. Most political crises are downstream of it. Understanding divergence structurally — rather than treating it as mere disagreement or moral failure — changes what kinds of interventions look plausible.

Its essential nature: a process by which value-bearing systems, under the influence of differential selection pressures, optimization dynamics, information environments, or internal feedback loops, move toward states that are mutually incompatible or incommensurable. The process is often irreversible beyond a threshold. It generates

coordination failures that cannot be resolved by appeal to a common framework, because the common framework is precisely what divergence has eroded.

---

## **PART II: HISTORICAL DEVELOPMENT**

### **Origins and Precursors:**

Human recognition of value divergence is as old as recorded encounter with the genuinely foreign. Herodotus documents the Persian Wars partly as a record of incompatible moral frameworks: Greek and Persian conceptions of legitimate governance, honor, and the sacred do not reduce to each other. Thucydides' Melian Dialogue is one of the earliest formal analyses of what happens when power asymmetry meets axiological distance — the Athenians and Melians are not merely disagreeing, they are operating from value structures that have diverged enough to make negotiation effectively impossible.

The Stoics tried to solve divergence by positing a universal logos that all rational beings participate in, generating a universal ethics. This move — positing a shared substrate to arrest apparent divergence — is repeated throughout intellectual history whenever divergence becomes politically urgent. Natural law theory in the medieval period, Kantian universalism in the Enlightenment, Rawlsian reflective equilibrium in the 20th century: each is, in part, an attempt to locate or construct a level at which diverging value systems converge.

The recognition that such convergence might be impossible, or might be a philosophical fiction, gains ground in the 19th century. Hegel's historicism locates values within the development of Spirit through historical forms, which implies that values are phase-specific: the values adequate to one historical moment are not adequate to another, and the transition generates genuine conflict rather than simple progress. Marx radicalizes this: values are class-specific, and apparent moral universalism is ideology — a local value system presenting itself as universal to serve local interests. Nietzsche goes further and frames all value systems as expressions of will to power, subject to genealogical decomposition, with no stable universal foundation. The implication is that divergence is the normal condition and convergence the artificial one.

Darwin had already delivered the biological version of this argument: species do not converge on a single optimal form. They diverge under differential selection pressure into forms that can no longer interbreed. The philosophical question that follows is whether value systems, like species, can diverge past the point of translation.

### **Developmental Phases:**

*Phase 1: Recognition without formalization (antiquity through the 19th century).* Divergence

is documented, managed, and sometimes theorized, but not yet treated as a structural problem requiring a formal solution. The dominant response is assimilation (convert the other to your framework) or separation (maintain distance). Contact zones between civilizations are managed through empire, trade, or war, not through value theory.

*Phase 2: The cultural relativism debates (late 19th to mid-20th century).* Anthropology makes divergence a scientific object. Franz Boas, Ruth Benedict, and Margaret Mead document that moral systems vary systematically across cultures in ways that resist hierarchical ordering. This generates a crisis in normative theory. If values are culturally relative, how do you adjudicate between cultures? The Holocaust forces the issue: moral relativism becomes politically untenable when confronting systematic extermination. The Universal Declaration of Human Rights (1948) is an attempt to arrest relativist divergence by fiat, positing a shared foundation the evidence for which remains contested.

*Phase 3: Value pluralism and incommensurability (mid-20th century).* Isaiah Berlin distinguishes pluralism from relativism: values can be genuine and non-reducible to a single framework without collapsing into the claim that all values are equally valid. Bernard Williams, Thomas Nagel, and others develop the incommensurability thesis: some values cannot be reduced to a common metric without losing what makes them valuable. Alasdair MacIntyre's *After Virtue* (1981) argues that modern moral discourse is a ruin of fragmented traditions, and that apparent moral disagreement is actually axiological divergence past the point of rational adjudication.

*Phase 4: Formalization in AI alignment (2000s to present).* The concept enters formal technical discourse. The problem of specifying a utility function for an AI that robustly tracks human values turns out to be the axiological divergence problem in a new register. Goodhart's Law (when a measure becomes a target, it ceases to be a good measure) is a weak version of it. The orthogonality thesis (intelligence and terminal goals are independent dimensions: any level of intelligence is compatible with almost any goal) implies that more capable systems will diverge further from human values if their goals were imperfectly specified. This is divergence under optimization pressure.

*Phase 5: Divergence across institutions, algorithms, and media (2010s to present).* Research on recommendation algorithms demonstrates that optimization for engagement produces value drift in users over time — gradual axiological divergence from prior states, driven by machine learning systems that have no understanding of the values they are reshaping. Political polarization research frames the widening of partisan value gaps as a divergence process accelerated by information environment fragmentation. These are divergence dynamics operating below the level of conscious choice.

### **Turning Points:**

The Holocaust as a limit case of what follows from divergence plus power asymmetry. The

Cold War as a sustained study in ideological divergence with mutual annihilation as the failure mode. The emergence of the internet as an accelerator of divergence through filter bubble formation. The AI alignment problem as a formalization of divergence that makes its structural character unavoidable.

### **Historical Myths:**

The myth of convergence: that economic development, education, and contact between cultures will cause value systems to converge on a liberal democratic baseline. The evidence for this is weaker than its proponents have claimed, and the mechanisms proposed (deliberative democracy, international institutions, free trade) presuppose enough axiological overlap to function, which is precisely what divergence erodes. A variant is the myth of rational consensus: that sufficiently sophisticated moral reasoning will reveal a shared foundation beneath apparent disagreement. Rawls's work is the most sophisticated attempt at this and remains contested.

---

## **PART III: STRUCTURAL AND SYSTEMIC ANALYSIS**

### **Core Components:**

Any instance of axiological divergence involves at least four elements: a set of agents or systems that are value-bearing; differential selection or optimization environments that exert pressure on those values; a feedback mechanism by which the agents' values influence their behavior, which in turn shapes their environment; and a prior state of relative axiological proximity from which divergence is measured.

### **Organizing Principles:**

Divergence is driven by optimization under different constraints. When two agents are rewarded (by biology, culture, institutional incentives, or explicit objective functions) for different behaviors, and when those behaviors are value-laden, their values will drift apart in proportion to the divergence of their reward landscapes. The deeper the optimization runs — the more thoroughly behavior is shaped by the reward landscape — the more complete the divergence becomes.

A secondary driver is self-reinforcement. Value systems are not passive. They generate behaviors that reshape environments, which select for further value drift in the same direction. A culture that values aggressive territorial expansion will tend to create conditions — internal hierarchies, resource dependencies, mythologies of chosenness — that further entrench and extend those values. A machine learning system optimizing for engagement will reshape its information environment in ways that make engagement optimization increasingly appropriate, narrowing the space in which other values can operate.

## **Feedback Loops:**

The most dangerous feedback loop in axiological divergence is the legitimacy loop: agents whose values have diverged tend to construct epistemic frameworks that render their own values self-evidently correct and the other's values confused or corrupt. This is not mere motivated reasoning. It is a structural feature of value systems under pressure: a value system that questioned its own foundations at every step would be paralyzed. The price of functional commitment is selective blindness to alternatives. The divergence then reinforces itself at the epistemic level, making detection and correction progressively harder.

## **Hierarchies:**

Divergence operates at multiple levels simultaneously: terminal values (what is intrinsically good), instrumental values (what produces good outcomes), and epistemic values (what counts as evidence and valid reasoning). Divergence at the epistemic level is the most dangerous because it removes the shared framework within which terminal and instrumental disagreements could in principle be adjudicated. When two parties cannot agree on what would count as evidence for a moral claim, argument is replaced by power.

## **Incentives:**

Institutional incentives are a major driver of divergence that tends to be underweighted in philosophical accounts. Organizations select for employees who share and advance their values, which means that over time, institutions diverge from the broader population's values while becoming more internally homogeneous. This is value divergence through institutional selection. Universities, media organizations, financial institutions, government agencies, and religious bodies all exhibit this dynamic. The result is a landscape of axiologically coherent but mutually divergent institutional cultures competing for authority to define public norms.

## **Contradictions:**

The deepest contradiction in the concept is this: the mechanisms by which divergence is detected and measured are themselves value-laden. To say that values have diverged is to imply a prior state of convergence that serves as a baseline. But that baseline was itself a historical artifact, a contingent configuration that was not axiologically neutral. The question of who chooses the baseline, and by what authority, is not a technical but a political one. Every measure of divergence encodes a preferred direction of convergence.

## **Dependencies:**

Axiological divergence is dependent on communication channels. Total isolation produces divergence of a kind but forecloses the kind of competitive divergence that generates real conflict. Contact without power symmetry tends to produce assimilation rather than divergence: the subordinate value system is suppressed or absorbed. The most productive

conditions for divergence involve contact with partial separation: enough interaction to generate comparative awareness, enough insulation to allow different optimization processes to run. Cold War ideological divergence is a case study.

### **Emergent Properties:**

Beyond a certain threshold of divergence, value systems become mutually illegible. This is not a failure of intelligence or goodwill. It is a structural consequence of divergence deep enough that the conceptual primitives of one system lack clean translations in the other. Medieval Catholic and modern secular humanist moral frameworks can gesture at each other across this gap, but neither can fully inhabit the other's logic. Emerging AI systems and human value systems may be heading toward a similar illegibility — not through malice but through differential optimization.

### **Failure Modes:**

The failure mode of managed pluralism is convergence theater: maintaining institutional frameworks of dialogue while the actual value distances continue to grow beneath them. The failure mode of divergence detection systems is Goodhart entrainment: the metrics used to measure divergence become targets, producing convergence on the metrics while actual divergence continues. The catastrophic failure mode is the elimination of one value system by another — through extinction, suppression, or assimilation — which resolves divergence by removing one of the diverging terms.

---

## **PART IV: PHILOSOPHICAL AND CONCEPTUAL ANALYSIS**

### **Ontology:**

Axiological divergence presupposes that values have some form of real existence: that they are not arbitrary labels but properties of agents and systems that have causal power over behavior. Whether values are taken as natural properties, social constructions, or evolutionary adaptations determines what kind of divergence is possible and what would count as resolving it. Moral realism implies that divergence is an epistemic phenomenon — agents have moved away from tracking moral truth. Anti-realism implies that divergence is the default condition, and convergence is a political achievement with no metaphysical guarantee.

The most interesting position is something like structural realism about values: values exist as functional properties of systems, real enough to have causal consequences, without needing to be grounded in a mind-independent moral order. On this view, divergence is a real process — it changes what systems do and what they can coordinate on — even if it does not involve departure from some external moral fact.

**Epistemology:**

Divergence raises a hard epistemological problem: from within a diverged value system, how would you recognize divergence? Each system has its own criteria for evaluating states, and those criteria tend to represent the system's own values as adequate to the situation. This is not pure dogmatism. It is a functional property of any sufficiently committed value system. The implication is that divergence tends to be detected from outside the diverging system, which means the detector must occupy a position of comparative access — a position that is itself value-laden. There is no view from nowhere in axiology.

This generates what might be called the divergence detection paradox: the agents whose values have diverged most are the least likely to recognize it, while the agents most capable of recognizing it are those who have maintained enough axiological distance to compare. Detection requires partially standing outside the system you are measuring.

**Human Nature:**

Every moral system embeds claims about what human beings fundamentally are and what they fundamentally need. When those claims diverge, the disagreement is not merely about what is good but about who we are. This is why axiological divergence between political ideologies tends to be so intractable: liberal and communitarian, individualist and collectivist framings differ not just in their prescriptions but in their models of the human being for whom those prescriptions are designed. Arguing across that gap without addressing the underlying anthropology is arguing about furniture while disagreeing about the architecture of the building.

**Power:**

Power enters axiological divergence at multiple points. Value systems with more institutional, economic, or military power can force apparent convergence — the defeated or subordinated party adopts the language of the dominant framework without necessarily adopting its logic. This produces a surface convergence that masks underlying divergence, which then reasserts itself when power asymmetry decreases. Imperial history is largely a record of this pattern. The values of subordinated peoples do not disappear under colonialism; they survive in transformed and sometimes subversive forms, re-emerging when the structural pressure lifts.

**Identity:**

Value systems are not merely held; they are constitutive of identity. To ask someone to abandon a core value is to ask them to become a different person, in a sense they often experience as annihilation. This is why axiological divergence generates resistance that looks disproportionate to the stakes as viewed from outside: the threat is not to a preference but to the coherence of a self. This dynamic is reproduced at the group level:



communities whose value systems are challenged experience the challenge as existential, because their collective identity is built from those values. Divergence that crosses a threshold of identity-constitution cannot be resolved by argument alone; it requires either time, shared experience, or external pressure sufficient to reshape identity.

### **Symbolism:**

Value systems express themselves symbolically, and symbols can maintain the appearance of shared values long after actual convergence has dissolved. Shared flags, sacred texts, constitutional language, and institutional rituals serve this function. They are interfaces between diverged subsystems, maintaining coordination on behavior without requiring agreement at the level of foundational values. The American Constitution functions this way: its language is shared; its interpretation has diverged. The symbol holds the system together while the divergence continues beneath it. When the symbol itself becomes contested — when different factions claim incompatible meanings for it — the buffer zone collapses and the underlying divergence becomes impossible to manage through normal institutional channels.

### **What Worldview Does Axiological Divergence Imply?**

Attending carefully to axiological divergence implies a worldview that is neither relativist nor universalist. It takes seriously that values are real and causally significant, that their configurations are historical, differential, and subject to structural pressure, and that convergence on a universal framework is a political and technical achievement — not a natural attractor state or a philosophical given. It is compatible with a form of moral realism that is chastened: moral facts may exist, but our access to them is mediated by value systems that are themselves subject to divergence. It is skeptical of any system that presents its own values as the natural endpoint toward which all reasonable beings converge, because that presentation is itself a divergence-management technique.

---

## **PART V: PSYCHOLOGICAL AND ANTHROPOLOGICAL DIMENSIONS**

### **Emotional Appeal:**

Axiological divergence as a concept serves different emotional functions for different audiences. For those experiencing the anxiety of moral fragmentation — the sense that shared normative ground is dissolving — it offers a structural diagnosis that is less immediately threatening than moral relativism. Divergence implies a prior convergence, which implies the possibility in principle of reconvergence. For those who experience themselves as representing an embattled minority value system, the concept legitimizes

their sense of difference: they are not simply wrong by the standards of the majority; they are differently optimized, the product of a different selection history. For those in AI alignment research, it converts what might feel like a speculative fear into a precise technical problem, which makes it tractable.

### **Archetypes:**

The mythic structure underlying axiological divergence is the fall from a state of primordial unity into differentiation and conflict. This narrative appears across cultures: Eden before the fall, the Tower of Babel (the moment when shared language — the medium of shared value — is shattered), the Mahabharata's account of dharmic fragmentation across the yugas, the Norse Ragnarok as a conflict among beings whose values have diverged past the point of peaceful coexistence. The persistence of this narrative structure suggests that the experience of value fragmentation is ancient and that cultures have consistently interpreted it as a fall from a better prior state.

Whether that prior state ever existed is a separate question. The Garden of Eden did not exist as a literal historical condition. But the narrative function is real: it provides a directionality (we have fallen) and a goal (return to unity), which makes divergence politically and religiously manageable. Secular political theory replaces the Garden with the Original Position, the State of Nature, or the Ideal Speech Situation: hypothetical constructs of prior convergence that legitimate current normative claims.

### **Rituals:**

Value systems maintain themselves partly through ritual — repeated practices that reinforce the values without requiring their conscious re-endorsement. Religious liturgy, national ceremony, legal procedure, and scientific peer review are all value-reinforcing rituals. They operate below the level of explicit axiological commitment, maintaining behavioral coherence within a value system while insulating it from the constant negotiation of its foundations. Axiological divergence, from a ritual-practice perspective, looks like desynchronization: the rituals of different groups no longer reinforce the same values, and the meaning of apparently shared rituals diverges.

### **Cognitive Biases:**

Confirmation bias, in-group favoritism, and motivated reasoning all accelerate axiological divergence by making it harder to detect. More relevant is what might be called value-attribution asymmetry: agents tend to attribute their own positions to values and their opponents' positions to interests. "We oppose this because it is wrong; they oppose it because it benefits them." This asymmetry, documented across political science research, converts what are genuinely value differences into moral condemnations, which escalates divergence by adding a layer of mutual delegitimization on top of the substantive value gap.

## **Memetic Dimensions:**

Value systems propagate through culture the way genes propagate through populations: with variation, selection, and heritability. Memes (in the original Dawkinsian sense) that encode values compete for host minds, and different information environments select for different memes. The fragmentation of media environments into self-reinforcing silos is a memetic selection pressure that accelerates axiological divergence: different meme pools optimize for different host value systems, which then become more differentiated and mutually illegible over time. Social media platforms are, among other things, divergence accelerators.

## **Why Are Humans Attracted to Axiological Divergence as a Concept?**

Because it names something they experience but lack precise language for. The sense that moral disagreement has become qualitatively different — that it is no longer about different conclusions from shared premises but about different logical systems with incompatible premises — is widespread. The concept also flatters the intelligence of its user: understanding divergence structurally, rather than blaming it on ignorance or bad faith, is a more sophisticated interpretive frame. And it points toward possible interventions (alignment, calibration, institutional design) rather than mere resignation, which satisfies the need for agency.

---

# **PART VI: ECONOMIC, POLITICAL, AND TECHNOLOGICAL FORCES**

## **Economic Incentives:**

Market economies generate axiological divergence through specialization and differential exposure to risk and reward. Financiers, agricultural workers, software engineers, and subsistence farmers occupy different positions in the economic structure, which subjects them to different selection pressures and produces different instrumental values. These instrumental differences accumulate into terminal value differences: different conceptions of fairness, different weights assigned to security vs. growth, different frameworks for assessing legitimate authority. The political science literature on economic voting is partly a record of axiological divergence driven by economic position.

The knowledge economy has intensified this. Cognitive labor, which produces abstract value, generates different value orientations than physical labor, which produces material value. The political divergence between urban professional and rural working-class populations in many countries is, at depth, an axiological divergence produced by differential participation in the knowledge economy.

## **Political Power Structures:**

Liberal democratic institutions are designed to manage axiological divergence — to produce binding collective decisions without requiring the resolution of underlying value conflicts. The separation of church and state is a divergence-management mechanism: by making religion private, it removes the highest-stakes value conflicts from the domain of state adjudication. Constitutional rights perform a similar function: they define the minimum conditions for participation regardless of divergent terminal values.

These mechanisms presuppose a degree of background convergence sufficient to make the framework itself legitimate. When divergence exceeds that threshold — when significant populations no longer endorse the framework's foundational values (pluralism, procedural fairness, individual rights) — the framework cannot manage the divergence it was designed to contain. Political polarization research suggests that advanced democracies may be approaching this threshold in several regions.

## **Technological Dependencies:**

Technology does not merely exist in a value-neutral environment. It co-evolves with values. Writing enabled the codification and transmission of value systems at scale, creating the conditions for large-scale convergence (world religions, empires, legal systems) and for precise articulation of divergence (heresy, apostasy, ideology). The printing press accelerated both: it spread Reformation theology (divergence from Rome) and eventually created the conditions for Enlightenment universalism (a new convergence proposal). The internet is producing a similar pattern at higher speed and lower legibility.

Recommendation algorithms deserve particular attention. They are axiological divergence machines: they optimize for engagement signals, which means they amplify the most emotionally activating value content, which tends to be the most extreme. Users who begin with moderate value orientations are gradually pulled toward positions that generate stronger engagement signals, while the algorithm's model of their preferences updates to reflect their new position. The result is divergence from prior states and from other users who were pulled in different directions by different engagement histories. This is not conspiracy; it is an emergent property of optimizing for a proxy metric (engagement) that is imperfectly correlated with the values the platform claims to serve (connection, information).

## **AI and Institutional Divergence:**

Large language models and other AI systems trained on different datasets, with different fine-tuning objectives, will exhibit axiological divergence from each other and from human value systems in ways that are difficult to measure. The models do not have values in the way humans do, but they have value-relevant tendencies — tendencies to represent certain options as better or worse, certain framings as more or less appropriate. As these systems

become more integrated into institutional decision-making, their tendencies will exert selection pressure on the value configurations of those institutions.

---

## **PART VII: RHETORICAL, SYMBOLIC, AND AESTHETIC ANALYSIS**

### **Language:**

The language of axiological divergence varies sharply by context. In AI alignment discourse, it is technical and formal: utility functions, objective misspecification, reward hacking, inner alignment vs. outer alignment. In political commentary, it becomes polarization, tribalism, culture war. In theology, it is heresy, schism, apostasy, heterodoxy. In anthropology, it is cultural relativity and incommensurability. Each vocabulary frames the phenomenon differently and implies different interventions, different sites of agency, and different distributions of blame.

The word “divergence” itself carries a geometrical neutrality — lines moving apart, no judgment attached — that is strategically useful in contexts where the parties do not agree on who has departed from the correct path. To say that values have diverged is to decline to say which side moved. This makes it useful as a diagnostic concept but potentially evasive as a normative one.

### **Branding:**

In contemporary political discourse, the concept of divergence tends to be branded differently by different political orientations. Conservative framings tend to locate divergence as departure from tradition: we have moved away from settled values. Progressive framings tend to locate divergence as failure of progress: we have not moved far enough toward justice. Both framings posit their own position as the baseline, which means both are, structurally, convergence demands dressed as divergence critiques.

The AI alignment community has done something more interesting: it has attempted to treat divergence as a technical problem separable from political value commitments, focusing on the formal properties of optimization processes rather than on which values are correct. Whether this technical frame is itself a political choice — favoring measurable over unmeasurable values, formal over tacit knowledge — is a legitimate critique.

### **Narratives:**

The dominant narratives around axiological divergence in contemporary culture are: the fall narrative (we were once unified and have fractured, implying a return is possible and desirable), the crisis narrative (divergence has reached a dangerous threshold requiring

urgent intervention), and the complexity narrative (divergence is a natural feature of complex adaptive systems, neither crisis nor tragedy). The fall narrative tends toward conservatism; the crisis narrative tends toward technocracy; the complexity narrative tends toward a kind of resigned systems thinking that often struggles to generate normative conclusions.

### **Artistic Representations:**

Literature has explored axiological divergence without naming it as such for millennia. The Iliad is largely a study in what happens when honor codes diverge within a coalition — Achilles and Agamemnon are not simply disagreeing about horses and girls; they have different and ultimately incompatible values around status, obligation, and the relationship between individuals and the collective project they serve. King Lear stages the divergence of value between generations. Dostoevsky's major novels stage divergence between religious and rationalist value systems with unusual evenhandedness.

In contemporary science fiction, axiological divergence is often the central problem: Greg Egan's *Diaspora* depicts posthuman clades whose values have diverged so far that they can barely communicate; Ursula Le Guin's Hainish cycle is partly a record of how value systems diverge under conditions of long isolation and what remains recoverable when contact is restored. These works treat divergence not as a problem to be solved but as a condition to be navigated.

---

## **PART VIII: CONTRADICTIONS, CRITIQUES, AND FAILURE MODES**

### **Internal Contradictions:**

The first major internal contradiction is the baseline problem. To measure divergence, you need a reference point — a prior state of relative convergence. But any such baseline was itself a historical artifact, the product of power structures, cultural hegemonies, and contingent events. Choosing the baseline implicitly endorses the value configuration of that moment. Nostalgia for a prior convergence that felt like unity often turns out to be nostalgia for a configuration in which one's own values happened to dominate. This is divergence-talk as legitimation.

The second contradiction concerns the normative ambiguity of convergence. If divergence is a problem, then convergence is the solution — but convergence toward what? Every convergence proposal installs some value system as the target. Proposals for universal human values, global liberal norms, AI value alignment, or international law all face the same challenge: the convergence they propose is convergence toward values that are themselves

contested. The call to converge is always, from some perspective, a call to submit.

The third contradiction is the detection paradox discussed earlier: the tools for detecting divergence are themselves embedded in value systems. The social scientist who measures political divergence is operating from a framework that treats certain research values (empirical evidence, replicability, democratic legitimation) as foundational. These are not neutral. They are a value system with its own divergence history and its own blind spots.

### **Ethical Concerns:**

Divergence-talk can naturalize what is political. If value differences are framed as the inevitable result of differential optimization pressures, this can obscure the fact that those pressures were themselves shaped by choices, power, and history. To say that rural working-class voters and urban professionals have axiologically diverged due to their positions in the economic structure is accurate as a structural description, but it risks treating as natural what was produced by specific policy choices, historical legacies of racial and economic exclusion, and deliberate decisions by institutions and corporations that shaped those structural positions.

### **Structural Weaknesses:**

The concept lacks a satisfying theory of resolution. Describing divergence in detail does not generate clear prescriptions for how to reverse it or stabilize it. Proposals for convergence through deliberation (Habermas), institutional design (Rawls), machine learning (AI alignment researchers), or cultural renewal (communitarians, traditionalists) each face the problem that they presuppose enough prior convergence to get started. If divergence has already consumed the common ground needed for deliberation to function, deliberative proposals are non-starters. If the training data for an AI value-learning system is already diverged, the AI learns divergence.

### **What Defenders Misunderstand:**

Defenders of the concept in the AI alignment context sometimes underestimate how much human value systems are themselves internally diverged and how much apparent convergence is a surface phenomenon maintained by shared vocabulary rather than shared meaning. "Human values" is not a stable target: it is a moving, contested, internally differentiated field. An AI aligned to human values as measured by some survey at some moment would be misaligned by the time the optimization ran.

### **What Critics Misunderstand:**

Critics who dismiss the AI alignment concern as science fiction or elite obsession sometimes miss that the formal structure of the problem — a powerful optimizer pursuing a specified goal that imperfectly tracks the intended goal — is not hypothetical. It is already operating in institutional settings, financial markets, recommendation systems, and hiring

algorithms, with measurable consequences for human welfare. The scale of harm is currently smaller than alignment researchers fear for future systems, but the structural pattern is already observable.

### **Strongest Critiques:**

The strongest critique of axiological divergence as a framework is that it tends to aestheticize intractability — to describe hard coordination problems in terms that make them feel structurally inevitable rather than politically contingent. This can discourage pragmatic interventions in favor of architectural solutions that are themselves politically freighted. The second strongest critique is that convergence on a single value framework — which is the implicit solution to a divergence problem — has historically been achieved through violence, suppression, and cultural destruction as often as through deliberation and mutual learning. The costs of convergence are as real as the costs of divergence.

---

## **PART IX: FUTURE TRAJECTORIES**

### **10-Year Trajectory (to approximately 2035):**

The dominant near-term pressure on axiological divergence is the continued development of AI systems that are trained on different datasets with different fine-tuning objectives for different institutional clients. These systems will exhibit measurable divergence from each other and from human value systems in ways that are not currently being systematically monitored. The lack of a shared measurement framework for AI axiology is itself a divergence-enabling condition.

Political polarization in liberal democracies will likely deepen, as the information environments driving divergence show no sign of reintegration. The institutional buffers designed to manage divergence — independent courts, professional norms, scientific consensus-formation — will continue to be contested by movements whose values have diverged from the liberal-proceduralist baseline those institutions presuppose.

The most significant near-term development may be the formalization of AI value alignment as a technical discipline with measurable standards. This would represent a genuine attempt to arrest divergence through engineered convergence, with all the political difficulties that implies.

### **25-Year Trajectory (to approximately 2050):**

At the 25-year horizon, the most significant variable is whether AI systems achieve a level of capability sufficient to exert strong selection pressure on human institutional and cultural value systems. A sufficiently capable AI system integrated into economic and governmental



decision-making would not merely reflect human values; it would begin to shape the environments that produce human values, creating a feedback loop. The question is not only whether the AI's values have diverged from human values, but whether the presence of the AI accelerates divergence among human value systems by fragmenting the shared environments in which common values are maintained.

The more pessimistic scenario involves the political fragmentation of currently unified legal and normative systems — international law, professional standards, scientific consensus — into competing blocs with incommensurable axiological foundations. This has happened before (the Cold War is a partial analogy) and resolved into renewed convergence under external pressure. The conditions for that kind of resolution in a multipolarity with AI systems embedded in each bloc's decision-making are not obvious.

### **100-Year Trajectory:**

At the century scale, two scenarios deserve attention. In the first, some form of axiological convergence is achieved through a combination of institutional design, AI value alignment research, and cultural pressure. This convergence would itself be a political achievement encoding specific values, and the suppression of divergent alternatives would be its cost. Historical analogies suggest this tends to require a crisis severe enough to make convergence feel like survival.

In the second scenario, divergence continues to the point where distinct value systems are maintained by distinct technical and institutional infrastructures — separate information environments, separate AI systems, separate governance frameworks — that do not share enough to coordinate on global-scale problems. This is not the end of humanity; it is a condition comparable to the pre-modern fragmentation of governance before the Westphalian system, but with higher-stakes coordination requirements (climate, pandemic, nuclear weapons, AI development). The cost of non-coordination is higher than it has ever been, which is the argument for treating convergence as urgent. The mechanisms for achieving it remain underdeveloped.

### **What Happens if Divergence Becomes Global Dominant Condition:**

The coordination failures multiply: climate treaties become unenforceable because parties' values regarding future generations, acceptable risk, and economic fairness have diverged beyond the point where compromise language holds. International institutions cease to function as genuine governance and become theaters of managed conflict. AI systems developed within diverged value systems produce outputs optimized for their local frameworks that are incompatible and potentially adversarial. The catastrophic failure mode is not a single event but an accumulation of unmanageable coordination failures.

### **What Happens if Divergence Disappears:**

Convergence to a single global value system would represent one of the most significant political achievements and one of the most significant losses in human history simultaneously. The loss is the permanent foreclosure of value alternatives that might have proven adaptive in conditions that have not yet arrived. Monocultures are fragile. A humanity that has converged on a single value framework has eliminated its axiological redundancy, which is a catastrophic risk in a world where the conditions that framework was designed for may change faster than the framework can adapt.

---

## PART X: DISTILLED INSIGHTS AND CONCEPTUAL TOOLS

**1. The Baseline Trap** Any claim that values have diverged from a prior state encodes a preference for that prior state. Recognizing this does not invalidate divergence analysis, but it requires treating the choice of baseline as a political act requiring justification, not a neutral technical decision.

**2. Divergence Scales with Optimization Depth** The deeper an optimization process runs — whether biological evolution, institutional selection, cultural reinforcement, or machine learning — the more complete the divergence it produces. Partial optimization produces partial divergence that remains negotiable. Total optimization produces agents whose values are no longer translatable across the optimization boundary.

**3. Epistemic Divergence Is More Dangerous Than Terminal Value Divergence** Two parties who disagree about what is good but share a framework for adjudicating that disagreement can in principle negotiate. Two parties who have diverged in their criteria for evidence and valid reasoning cannot. The former is a political problem; the latter is a civilizational one. Monitoring divergence at the epistemic level is more important and harder than monitoring it at the normative level.

**4. Symbols Are Divergence Buffers** Shared symbols — constitutional language, religious texts, scientific norms, institutional rituals — maintain behavioral coordination despite underlying axiological divergence. When the symbols themselves become contested, the divergence they were masking becomes immediately politically active. This is why symbol disputes (flag controversies, textbook battles, monument removals) generate heat disproportionate to their apparent stakes: they are not about the symbols but about the divergence the symbols were containing.

**5. Convergence Is a Political Achievement, Divergence Is a Physical Default** Left to their own selection pressures, value systems will differentiate. Convergence requires active institutional, cultural, and technical maintenance. This inverts the folk assumption that agreement is natural and conflict is aberrant. It implies that the appropriate response to divergence is sustained investment in convergence-maintaining institutions, not optimism

that disagreement will naturally resolve.

**6. The Specification Problem Is Universal** Goodhart's Law applied to values: any specification of what you value is an imperfect proxy for what you actually value, and optimizing the specification hard enough will produce behavior that maximizes the proxy while undermining the underlying value. This applies to constitutions, moral codes, AI objective functions, and institutional mandates. The implication is that robust value systems require ongoing calibration and the preservation of interpretive flexibility — which is itself a value claim that conflicts with the desire for precision.

**7. Illegibility as a Phase Transition** Axiological divergence is not linear. It has phase transitions: levels of divergence at which translation becomes functionally impossible, even with good faith and high intelligence on both sides. Detecting proximity to a phase transition is a critical capability. Beyond it, the tools of deliberative democracy, international negotiation, and rational discourse are not merely strained; they are structurally inadequate to the problem.

**8. Value Systems Are Ecologies** A landscape of multiple diverged value systems is an axiological ecology. Ecologies have resilience properties that monocultures lack. This is an argument against convergence as a terminal goal: some degree of axiological diversity is an adaptive property of civilizational systems, a hedge against the possibility that the currently dominant value configuration is wrong in ways that are not yet apparent.

**9. The Divergence of Divergence-Management** The institutions, disciplines, and technologies designed to manage axiological divergence are themselves subject to it. Philosophy departments, international law, AI alignment research, and political mediation are all internally diverged and embedded in political contexts that shape their outputs. There is no neutral ground from which to manage divergence. Management always encodes values.

**10. Timing Is Everything** Divergence that is detected and addressed early is reversible. Divergence that has run long enough to generate self-reinforcing epistemic frameworks, institutional selection effects, and identity-constituting commitments approaches irreversibility. The window for convergence is real but finite. Recognizing divergence as a process rather than a state implies that the time at which intervention is applied is as important as the nature of the intervention.

---

## PART XI: FINAL SYNTHESIS

Beneath the technical vocabulary of utility functions, phase transitions, and feedback loops, axiological divergence is an old story. It is what happens when the conditions that produced a shared world change faster than the world can update. It is the gap between what a map

says and what the territory has become. It is the moment when people who once shared a language discover they have been using the same words for different things for long enough that the difference has become structural.

What it reveals about humanity is uncomfortable. Shared values are not the natural state from which conflict is a deviation. They are a managed condition, requiring continuous maintenance, institutional support, deliberate communication, and enough shared material reality to give the maintenance something to grip. Remove those conditions — through economic divergence, information fragmentation, geographic isolation, or the rapid introduction of optimization systems that bypass the channels through which shared values are formed and reinforced — and the values diverge. The process is not dramatic. It is quiet. By the time it is visible, it has usually run long enough to be difficult to reverse.

What humanity fears in axiological divergence is not conflict per se. Conflict between parties who share enough framework to negotiate is manageable. What is feared — what Babel mythologizes and AI alignment researchers formalize — is the loss of the shared language in which disputes can even be articulated. A world in which value systems have diverged past mutual legibility is not a world of continuous war; it is a world in which coordination on the problems that require coordination becomes structurally impossible, and the costs accumulate silently until they are catastrophic.

What it promises, in its more complex readings, is also real. A landscape of diverse value systems is an exploration of possibility space. Not all experiments succeed, and the unsuccessful ones impose costs on those who live inside them. But the alternative — a civilization that has converged on a single value framework and ceased to explore alternatives — has its own fragility, which history documents with regularity. Rome converges; Rome falls. The Church converges; the Reformation fragments it. Liberal democracy converges in the mid-20th century; the pressures of the 21st reopen its fault lines. Convergence is not a final destination; it is a phase.

The contradictions that define axiological divergence are not resolvable. You cannot want genuine value pluralism and also want a stable common ground for coordination. You cannot want precise AI value alignment and also want the values to be living, contested, and capable of revision. You cannot want institutional robustness and also want institutions that are responsive to axiological change. These tensions are not failures of thinking. They are features of the problem.

What divergence points toward — the future it makes necessary — is a politics and a technology of divergence management sophisticated enough to maintain coordination capacity without requiring convergence, and wise enough to preserve axiological diversity without sacrificing the minimum common ground on which survival-level coordination depends. That is a harder problem than either enforced uniformity or resigned pluralism. It is the problem that the 21st century has handed civilization without providing the tools to

solve it.

The myths surrounding it — Eden, Babel, the Fall — suggest that humans have always known this problem and have always reached for metaphysical solutions to what is, at depth, a practical and political one. The metaphysical solutions keep failing. The practical ones keep being deferred. The gap between what we know about divergence and what we are willing to do about it is itself a kind of axiological divergence: a distance between what we value in theory and what we optimize for in practice.

That gap is where the work is.

---

*Analysis produced under the Neil Protocol Engine / NBMHouse Research Series*  
*Nox Vale — May 2026*